Producing small area estimation using R in the Romanian official statistics

Authors:

Ana Maria DOBRE*
Nicoleta CARAGEA**

Abstract: The purpose of this paper is to reveal the opportunities found in the Romanian official statistics to develop a long-way but strong implementation of the Small Area Estimation techniques, together with the use of R statistical software. The Small Area Estimation technique, a model-based approach to produce regional or even locality level data, has been already succesfully applied on the estimation of the international migration, which is often hard to estimate.

In these circumstances, the official statistics in Romania face new challenges in data analysis tools. In the last two years, a small team of statisticians introduced R both in the official statistics and academia, as an opportunity for progress.

The paper contains an overview of the small area techniques applied to the estimation of international migration and covers other possible applications of small area estimation methods, presenting the challenges currently considered in the official statistics.

Keywords: Small Area Estimation, Labour Force Survey, R Software, Official Statistics, International Migration

JEL Classification: C13; C51; C88; F22

---

* *Ana Maria DOBRE, National Institute of Statistics & Institute of National Economy, Romanian Academy, e-mail: dobre.anamaria@hotmail.com*

** *Nicoleta CARAGEA, National Institute of Statistics & Ecological University of Bucharest, Faculty of Economics, e-mail: nicoletacaragea@gmail.com*

1. Introduction

The lack of exact figures on emigration led to the need for a new statistical thinking based on econometric models.

In the context given by the difficulty in measuring the phenomenon of international migration, the official statistics in Romania developed and implemented a methodology based on Small Area Estimation (SAE) techniques.

In the large-scale phenomenon of international migration, estimation for specified regions or areas are needed. The estimates from small areas are obtained by assuming that they are similar to the entire population.

The computing solution for applying Small Area Estimation techniques was R, the most powerful data analysis tool. R has been choosen since it is by far the most used open source statistical software among data scientists and academic communities. The statistical software R shows the advantages of an open source system: low costs (the cost of using R are related only with training of users), easy customization and use of packages, technical support provided by a large community of users, continuous upgrade.

2. Literature review

Back in the 80s, Purcell and Kish (1980) described the estimation procedures for small areas using census data and aditional information from administrative data or similar sources. An extenstion of their view was created by Ghosh and Rao (1994).

Nevertheless, some important scientific approaches of Small Area Estimation are those of Rao (2003) and Sarndall (1984). These methods have been implemented by Breindanbach (2011) in the *JoSAE* package in R. Breidenbach (2012) developed Small Area Estimation methods for forest attributes, but further scientific requierements showed that Small Area Estimation methods could be used on any gathering issue.

Especially in the last decade, the interest for Small Area Estimation techniques in official statistics was really increasing. Several projects were successfully accomplished, having as main purpose the study, development and implementation of Small Area Estimation in European statistics offices. The Office for National Statistics in the United Kingdom ran a large project named EURAREA 2001-2003, a project for enhancing Small Area Estimation techniques to meet European needs with participating countries from the European Union. Another important project was SAMPLE 2009-2011 (Small Area Methods for Poverty and Living Condition Estimates), which aimed to develop efficient and reliable indicators

and high quality data on life conditions not only at national level but also at small area levels. The European Commision supported another project aimed to cover the topic of Small Area Estimation on Poverty Statistics – AMELI 2011 (Advanced Methodology for European Laeken Indicators). Afterwards, another European project was ESSnet for Small Area Estimation. Its general objective was to develop a framework enabling the production of small area estimates for ESS social surveys.

3. The methodology based on Small Area Estimation and R

Small Area Estimation seeks to improve the precision of the estimates when standard methods are not accurate enough. Thus, Small Area Estimation method produces estimations for the areas having not reliable direct estimators. The conceptualization of this method is somehow confusing, because this technique does not require that the areas be small, but the number of statistical units selected from the respective areas must be low. Also, the areas are sometimes called domains, due to the fact that they do not have to be geographical, but could refer to cross-classified demographic categories or other kind of categories, like a group of economic activities.

In Romania, the applied methodology of Small Area Etimation is used to produce data on annual international migration outflows. In order to meet the requirements of the SAE procedure, two data sources are combined: the Labour Force Survey (LFS), a statistical survey containing the variable of interest and a set of covariates and the 2011 Population and Households Census containing the same covariates and an auxiliary variable (Resident population). The common variables of LFS and Census were used to identify the possible correlates of migration incidence.

Thus, the following variables are used in the model:

- Binary indicator variable for migration (survey variable): *absent*
- Covariates
  - Gender: male and female;
  - Age group: age in five categories (age 0-14, age 15-24, age 25-39, age 40-64, age 65 and over);
  - Residence area: urban and rural;
  - Education level: education in three categories (low, medium and high);

- Activity: activity in five categories (employed, unemployed, pupil/student, pensioner and other situations);
- Marital status: marital status in three categories (single, married, widow/divorced).

LFS migration data consists of binary variables at the unit level. The variable of interest for the model is *absent* explained as it follows:

- 1 – If people are absent from home for more than 12 months;
- 0 – Otherwise.

Hereby, the categorical effect is represented by the migrant/non-migrant state of the person.

Based on combined data from the both sources, the application of the Small Area Estimation model makes possible to estimate the stock of emmigrants at county level.

The model is explained schematically in the Appendix, based on an exemplification of one of the covariates – age group. The LFS structure contains the migration status by age group, at unit level (individual data). The Census data are structured by small areas and age group. The specific variable of LFS is "Migration status" and the specific variable of Census data is "Resident population". By combining both data sources, estimates at small area level will result. Actually, different forms of the procedure could be developed depending on the completeness of the information available in both data sources.

The computation method used was *JoSAE* package and *nlme* package of the R software.

The model for estimating migration is based on linear regression models with mixed effects (fixed or random), hereby *lme* function (Linear mixed-effect model) is used**.** The regression coefficients are computed by maximum likelihood – Restricted Maximum Likelihood (REML). Below, we present the *lme* function in R, its output and the explanations in Table 1, applied on a simulation on small number of individuals:

```
> summary(fit.lme <- lme(absent ~ edu_med + edu_high + activ3 + activ4 + age_2 + age_3, data = LFS, random = ~1 | county))
```

Table 1. The output of the lme function in R

Output	Explanations
Linear mixed-effects model fit by REML Data: LFS AIC BIC logLik -1013620 -1013517 506819	AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) criteria that can be used to select the model (for example, for two different models, a lower AIC denotes a better model)
Random effects: Formula: ~1 | county (Intercept) Residual StdDev: 0.027727 0.112331	Estimates of dispersion parameters – random effect
Fixed effects: absent ~ edu_med + edu_high + activ3 + activ4 + age_2 + age_3 Value Std.Error DF t-value p-value (Intercept) 0.011774957 0.0006298361 661511 18.69527 0.0000 edu_med 0.009224348 0.0003336516 661511 27.64665 0.0000 edu_high -0.000901400 0.0007034765 661511 -1.28135 0.2001 activ3 -0.021757633 0.0004961657 661511 -43.85155 0.0000 activ4 -0.013458106 0.0003602981 661511 -37.35269 0.0000 age_2 0.022043679 0.0004960532 661511 44.43814 0.0000 age_3 0.022883402 0.0004074020 661511 56.16909 0.0000 Correlation: (Intr) edu_med edu_sup activ3 activ4 age_2 age_3 edu_med -0.224 edu_high -0.103 0.309 activ3 -0.162 0.270 0.140 activ4 -0.265 0.220 0.121 0.251 age_2 -0.071 -0.103 -0.036 -0.378 0.147 age_3 -0.145 -0.122 -0.129 0.113 0.292 0.172	Estimated values of the parameters; Standard deviations; T-student significance test; Correlation matrix.
Standardized Within-Group Residuals: Min Q1 Med Q3 Max -1.80565643 -0.22612616 -0.05270228 0.04283191 8.94436024 Number of Observations: 664703 Number of Groups: 42	Descriptive statistics of residual values; Number of observations; Number of groups.

Source: Own computations in R.

The linear mixed regression models were applied because the individual data (unit level) were organized into clusters (area level). The theoretical values of the variable of interest are correlated with covariates through a regression function. In general, the regression parameters can be generated by fixed effect regression (the number of coefficients is equal to the number of covariates), or could be generated by random effect (the number of coefficients is multiplied by the number of areas).

In order to ensure representativeness by small areas, the estimators must be unbiased; therefore the estimated average of the variable of interest must represent all the statistical units in the sample. Unbiased estimators are usually obtained by selecting very large samples, the selection comprising statistical units distributed in all the small domains (design-unbiased estimators). In this case direct estimators can be used successfully.

But in our case it was not possible to use direct estimates because they were not reliable. Therefore, it was necessary to use econometric methods to compute unbiased estimators (model-unbiased estimators).

The JoSAE package in R produces three types of estimators: GREG estimator, Synth estimator and EBLUP (Empirical Best Linear Unbiased Prediction) estimator.

The GREG estimator is obtained by adjusting the direct estimator Horwitz-Thomson with the differences between auxiliary variables environments calculated for each area of both data sources, Labour Force Survey and Census.

The Synthetic estimator involves linearity between the variable of interest and covariates for all areas, including those that were not included in the sample.

The model based-estimates and the design-based estimates are combined to produce EBLUP estimates. In statistics, BLUP is the resulting value of a predictor based on linear mixed models to estimate parameters due to regression's random effects. We used EBLUP estimators as final value for the estimates for a better accuracy. They are said to have the best properties of both design-based and model-based estimates.

The Synth, GREG and EBLUP estimates are obtained by measures of a single function of JoSAE package:

```
> result <- eblup.mse.f.wrap(domain.data = d.data, lme.obj = fit.lme)
```

The function *eblup.mse.f.wrap* computes estimates at county level as shown in Table 2.

Table 2. The types of estimators produced by SAE method in R

County	EBLUP	GREG	Synth
1	37660	37624	37624
2	47660	47524	47524
3	64660	64624	64624
4	63660	66324	66324

Source: own computations.

The eblup.mse.f.wrap function also computes the variances of the EBLUP, GREG and Synth estimates. The wrap function returns a data frame with many entries for every domain:

- Predictor variables obtained from the domain data;
- Mean of the predictor variables and response observed at the samples;
- Number of samples;
- Mean residuals of a linear model and the linear mixed-effect model;
- EBLUP, GREG and Synth estimates of the mean of the variable of interest;
- Variance of the means for the sample;
- Components of the EBLUP variance and the standard errors derived from the variances.

A huge challenge is the accurate choice of the most plausible estimation model and this is essential by compromising between minimization of the variance of estimators and their displacement beside the mean values. Also, auxiliary variables used must be correlated with the variable of interest and the correlations between variables must be strong. However, the selection of auxiliary variables depends on data availability (on their existence in the small area level). Along with small area estimates we computed model selection measures: Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC).

Small Area Estimation techniques were used to obtain estimates at county level (NUTS 3), with reliable coefficients of variation with values between 0.88 and 3.94. The minimum value 0,88 shows a high-level of emigration in the given county, meanwhile the maximum value of 3.94 shows a low level of migration. In

Table 3 we present the distribution of the coefficients of variation in the 42 counties:

Table 3. Distribution of the coefficient of variation of the SAE estimates

Coefficient of variation	Frequency	Percentage
0.00-1.00	2	4.87
1.01-2.00	21	51.21
2.01-3.00	12	28.25
3.01-4.00	7	15.67

Source: own computations.

Also, just to ensure the accuracy of the estimations, the benchmark method was used, comparing the results of the Small Area estimation model with the mirror statistics from Eurostat Migration Data.

Actually, the model is of course perfectible; there are some challenges like the non-migration character of Labour Force Survey, the data availability or the imposibility to estimate on disaggregation levels.

4. Further research opportunities

Small Area Estimation techniques could be successfully applied to various social and economic statistics fields.

A special attention is focused on the poverty measurement. One of the conclusion exposed by UNECE (2013) of the in-depth review of poverty statistics was that «an important direction for developing poverty statistics is to improve the reliability of poverty measurement on the basis of indirect assessment methods, including for small areas.».

The model could be applied for producing estimates of poor individuals in small areas, such as county level. Also, using EU-SILC, Census data, or other statistical and administrative sources the SAE method could be used for estimating either the number of people at risk of poverty or social exclusion, poor persons, persons living in households with low work intensity, or individuals facing severe material deprivation. The World Bank has already used SAE methods for shaping the Poverty Maps and Shared Prosperity and they have published their results in Bedi's study (2007). One of the beneficiary countries was Romania (2013).

Another possible application of Small Area Estimation techniques is the small area estimates of labour status or educational level, based on Labour Force Survey.

The estimates obtained by Small Area Estimation Techniques would contribute to policy efforts aimed at reducing poverty, inequality and social exclusion, helping Member States in general and Romania in particular to progress towards the goals of the Europe 2020 Strategy or to design better policies.

Acknowledgment

This paper has been financially supported within the project entitled „*SOCERT. Knowledge society, dynamism through research*", contract number POSDRU/ 59/1.5/S/132406. Thisproject is co-financed by European Social Fund through Sectoral Operational Programme for Human Resources Development 2007-2013. *Investing in people!*

References

Bedi, T., Coudouel, A., Simler, K. (2007), "More Than a Pretty Picture – Using Poverty Maps to Design Better Policies and Interventions", The World Bank, available at: http://siteresources.worldbank.org/INTPGI/Resources/342674-1092157888460/493860-1192739384563/More_Than_a_Pretty_Picture_ebook.pdf.

Breidenbach, J. (2011), JoSAE: "Functions for unit-level small area estimators and their variances. R package version 0.2.", http://CRAN.R-project.org/package=JoSAE.

Breidenbach, J., Astrup, R. (2012), "Small area estimation of forest attributes in the Norwegian National Forest Inventory". *European Journal of Forest Research*, 131, 1255-1267.

Caragea, N., Alexandru, A.C., Dobre, A.M. (2012), "Bringing New Opportunities to Develop Statistical Software and Data Analysis Tools in Romania", *The Proceedings of the 6th International Conference on Globalization and Higher Education in Economics and Business Administration*, ISBN: 978-973-703-766-4.

Dobre, A.M., Caragea, N., Alexandru, C. (2013), "R versus Other Statistical Software*", Ovidius University Annals*, 13, 484-488.

European Comission, "Advanced Methodology for European Laeken Indicators", https://www.uni-trier.de/index.php?id=25157&L=2.

European Comission, "Small Area Methods for Poverty and Living Condition Estimates Project", http://www.sample-project.eu/.

Ghosh, M., Rao, J.N.K. (1994), "Small area estimation: an appraisal". *Statistical Science*, 9, 55-93

Istituto Nazionale di Statistica, Italy, "ESSnet for Small Area Estimation", http://www.cros-portal.eu/content/sae.

Office for National Statistics (ONS), United Kingdom, "EURAREA Project", http://www.cros-portal.eu/content/eurarea.

Pinheiro J., Bates D., DebRoy S., Sarkar D., R Core Team (2014), nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-115, http://CRAN.R-project.org/package=nlme.

Purcell, N. J., Kish, L. (1980) "Postcensal estimates for local areas (or domains)", *International Statistical Review*, 48, 3-18.

R Development Core Team, 2005, "R: A language and environment for statistical computing". R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org.

Rao, J.N.K. (2003), *Small Area Estimation*, John Wiley & Sons, Hoboken, New Jersey.

Rao, J.N.K., Sinha, S.K. (2008), "Robust Small Area Estimation under Unit Level Models", *Proceedings of the Survey Research Section*, American Statistical Association, 145-153.

Rao, Poduri S. R. S. (2000) *Sampling Methodologies with Applications*, Chapman and Hall/CRC Press, London and New York.

Sarndal, C. (1984), "Design-consistent versus model-dependent estimation for small domains", *Journal of the American Statistical Association*, JSTOR, 624-631.

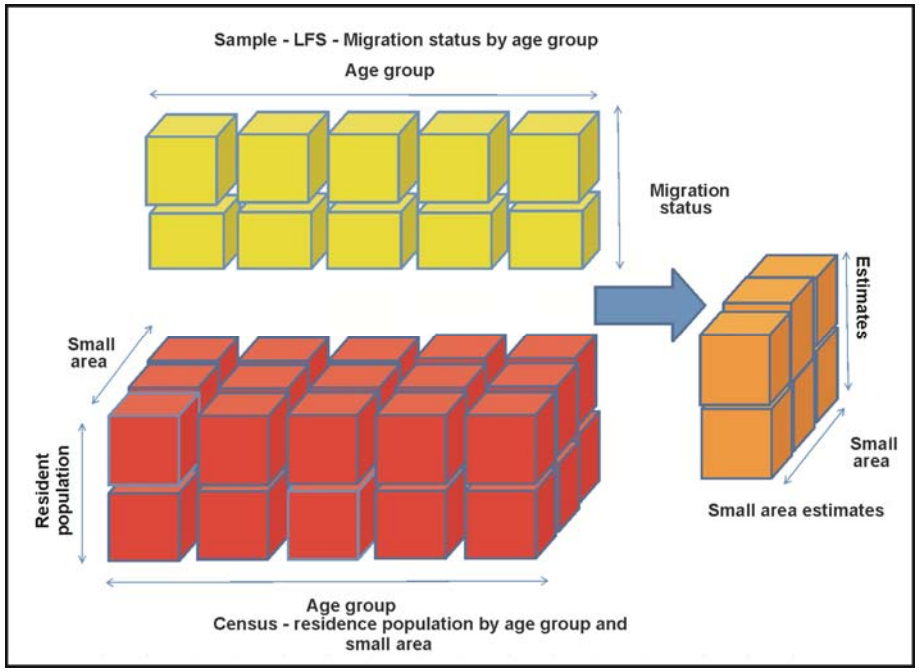
Statistics Division of United Nations (2002), "Tehnical Report on Collection of Economic Characteristics in Population Censuses", http://unstats.un.org/unsd/demographic/sconcerns/econchar/Series_M_119.pdf.

Vergil. V., Caragea, N., Pisica, S. (2013), "Estimating International Migration on the Base of Small Area Techniques", *Journal of Economic Computation and Economic Cybernetics Studies and Research*, Bucharest, 3, http://www.ecocyb.ase.ro/nr.3.pdf/Voineagu%20Vergil.pdf

United Nations Commision for Europe, Conference of European Statisticians 2013, http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2013/7.pdf.

Appendix

The procedure of Small Are estimation model on international migration



Source: Authors' view, adapted after the Tehnical Report on Collection of Economic Characteristics in Population Censuses, Statistics Division of United Nations (2002)